Mining Flow Cytometry Data

Sofie Van Gassen Sofie.VanGassen@irc.vib-UGent.be
Celine Vens Celine.Vens@irc.vib-UGent.be
Tom Dhaene Tom.Dhaene@intec.ugent.be
Yvan Saeys Yvan.Saeys@irc.vib-UGent.be
VIB Inflammation Research Center and Department of Respiratory Medicine, Ghent University
Technologiepark 927, 9052 Gent, Belgium

Department of Information Technology (INTEC)-iMinds, Ghent University
Gaston Crommenlaan 8 (Bus 201), 9050 Gent, Belgium

Keywords: data mining, application, flow cytometry, clustering

We want to introduce flow cytometry data mining to the Benelearn community. Flow cytometry (Herzenberg et al., 2006; Aghaeepour et al., 2013) allows to quantify different cell populations in large numbers of cells, by suspending the cells in a stream of fluid and passing them through a laser beam, while measuring the resulting fluorescent and scattered light. Flow cytometry is applied both in clinical settings (in the diagnosis of health disorders) and in research settings. For every input sample, 10,000 up to 1,000,000 cells are measured, each described by 10 up to 20 features. However, recent technological advances in flow cytometry result in techniques that will be able to measure up to 100 dimensions. Standard practice is that pathologists or researchers manually detect different cell populations in the sample by iteratively identifying regions of interest in two-dimensional scatter plots, a process that is labor-intensive and subjective. In this talk we will discuss challenges and opportunities for automated analysis of flow cytometry data.

One of the first steps in the analysis of this type of data is the use of clustering techniques to identify populations of cells in an objective and reproducible way. Important challenges for clustering these datasets include: populations with different densities; populations that are not elliptic-shaped; rare populations that can be easily confused with noise; hierarchical populations that consist of several sub-types;... Another challenge is that cells can be in transition from one population to another, and thus can belong to several clusters at the same time. As individual patient samples may consist of tens of thousands up to millions of cells, and clinical datasets often consist of hundreds of patient samples, scalability is an important aspect as well. Finally, since flow cytometry labs typically have large amounts of manually clustered samples available, this information should be incorporated into the clustering process, e.g. by using constraint-based clustering or transfer learning techniques.

In addition, often other types of metadata, such as clinical information, is available for the samples, in addition to flow cytometry data. In this context, we obtain a relational dataset of samples (patients), where each sample is related to a large set of cells. The most natural way for relational learning techniques to summarize the set of cells is to aggregate over clusters of cells of the same population. This involves learning a similar clustering structure for every sample: patients have the same cell types, although the numbers may be highly affected by a disease being present or not.

Acknowledgments

Sofie Van Gassen is funded by a Ph.D. grant of the Agency for Innovation by Science and Technology (IWT).
Celine Vens is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO).

References

Aghaeepour, N., Finak, G., FlowCAP Consortium, DREAM Consortium, Hoos, H., Mosmann, T., Brinkman, R., Gottardo, R., & Scheuermann, R. (2013). Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods*, *10*, 228–238.

Herzenberg, L., Tung, J., Moore, W., Herzenberg, L., & Parks, D. (2006). Interpreting flow cytometry data: a guide for the perplexed. *Nature Immunology*, *7*, 681–685.